Exploring the Web of Data for Earth and Environmental Sciences

Xiaogang (Marshall) Ma

Tetherless World Constellation
Rensselaer Polytechnic Institute

max7@rpi.edu rpi.edu/~max7 @MarshallXMa

x.marshall.ma MarshallXMa 0000-0002-9110-7369

Outline

- Web of Data
  - Semantic Web, RDF, Ontology, Linked Open Data
- Weaving the Web of Data
  - OneGeology-Europe
  - Global Change Information System
  - Deep Carbon Observatory-Data Science
  - Deep Time Data Infrastructure
- Exploring the Web of Data
  - Semantic similarity
  - Concept mapping
- Summary
  - Semantic eScience

Semantic Web

"The **Semantic Web** is an extension of the current web in which information is given well-defined **meaning**, better enabling computers and people to **work in cooperation**. "

Berners-Lee et al., 2001. *Sci. Amer.*

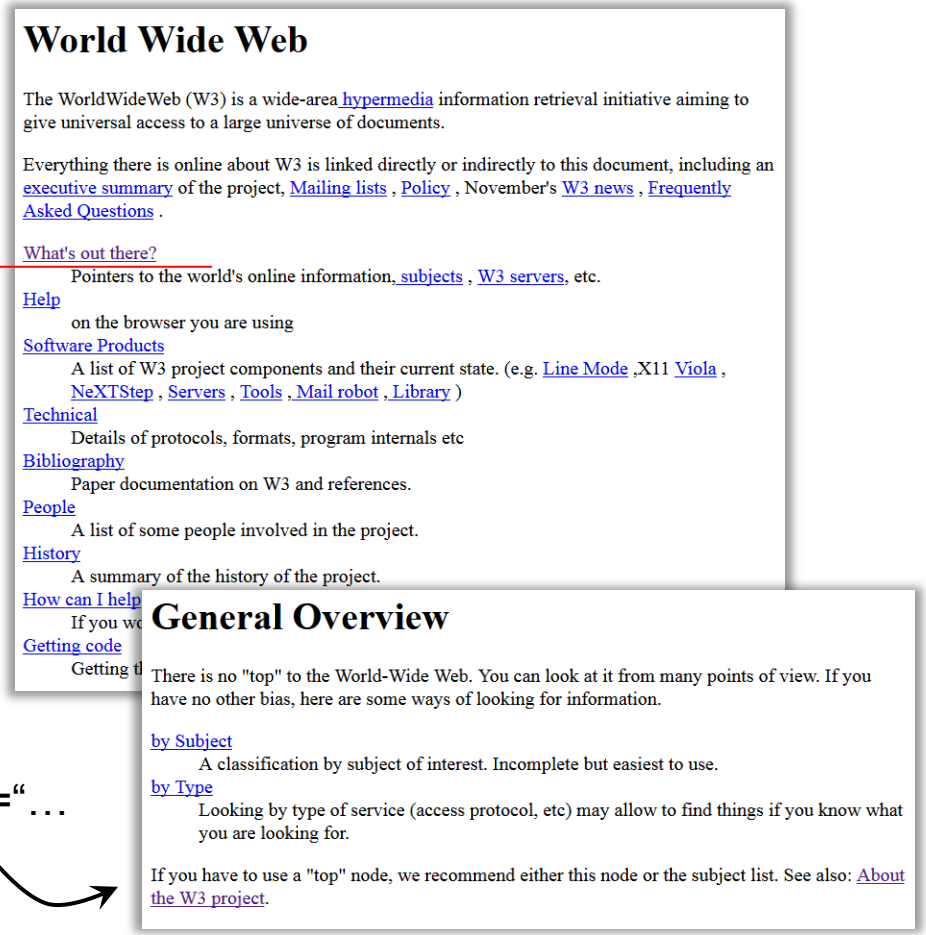
Web of Documents vs. Web of Data

Back to the early 1990s

- HTML and URL
- Markup language and ways for connecting resources
- Below the file level
- Stopped at the text level

The First Website

http://info.cern.ch/hypertext/WWW/TheProject.html



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help
If you w
Getting code
Getting t

General Overview

There is no "top" to the World-Wide Web. You can look at it from many points of view. If you have no other bias, here are some ways of looking for information.

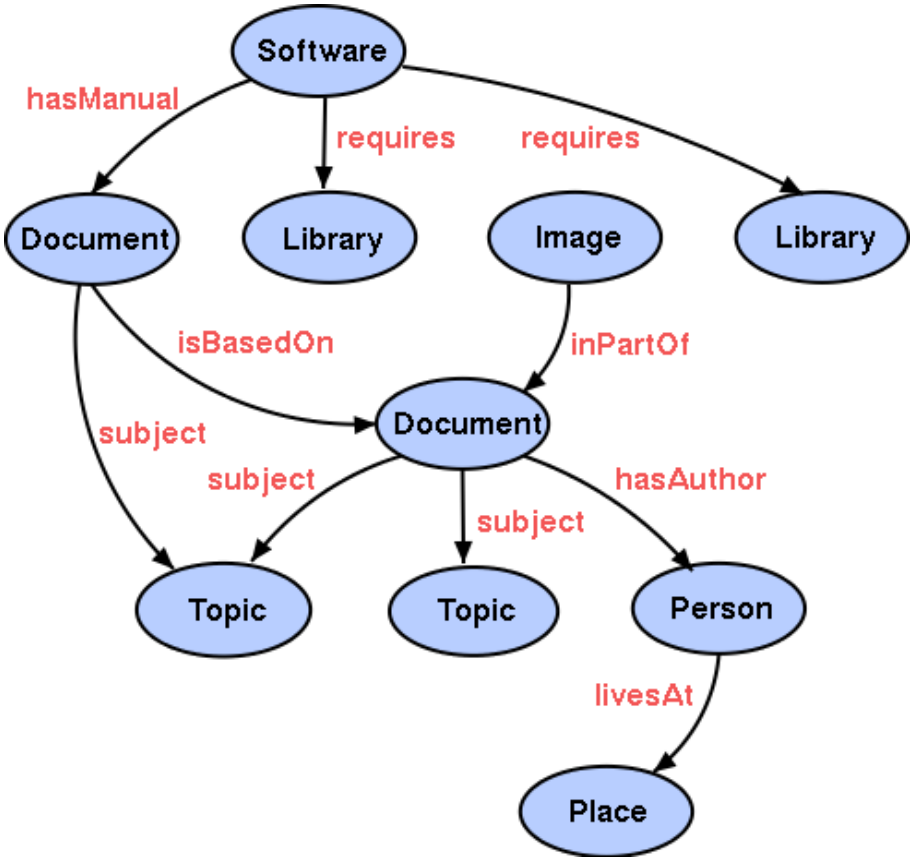
by Subject
A classification by subject of interest. Incomplete but easiest to use.
by Type
Looking by type of service (access protocol, etc) may allow to find things if you know what you are looking for.

If you have to use a "top" node, we recommend either this node or the subject list. See also: About the W3 project.

Web of Documents vs. Web of Data (cont.)

Since the early 2000s…

- XML, RDF, OWL and URIs
- Markup language and ways for connecting resources
- Below the file level
- Below the text level
- At the data level



Miller, 2004

Resource Description Framework (RDF)

- A standard of W3C
- RDF is made up of triples
  - <**s**ubject, **p**redicate, **o**bject>

<Mozart, composed, The Magic Flute>
<Mozart, isA, Musician>
<The Magic Flute, isA, Opera>

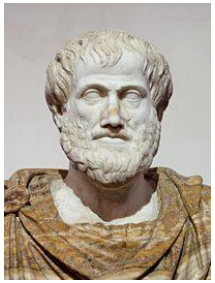
- RDFS extends RDF with a standard “ontology vocabulary”
  - Class, Property
  - subClassOf
  - Domain, Range
  - Type
  - …

<Musician, rdf:type, owl:Class>
<Musician, rdfs:subClassOf, Artist>
<composed, rdf:type, owl:ObjectProperty>
<composed, rdfs:domain, Musician>
<composed, rdfs:range, Opera>

http://www.w3.org/RDF/ http://www.w3.org/TR/rdf-schema/

Ontology

- The term ontology is originated from philosophy
  - The study of the nature of existence

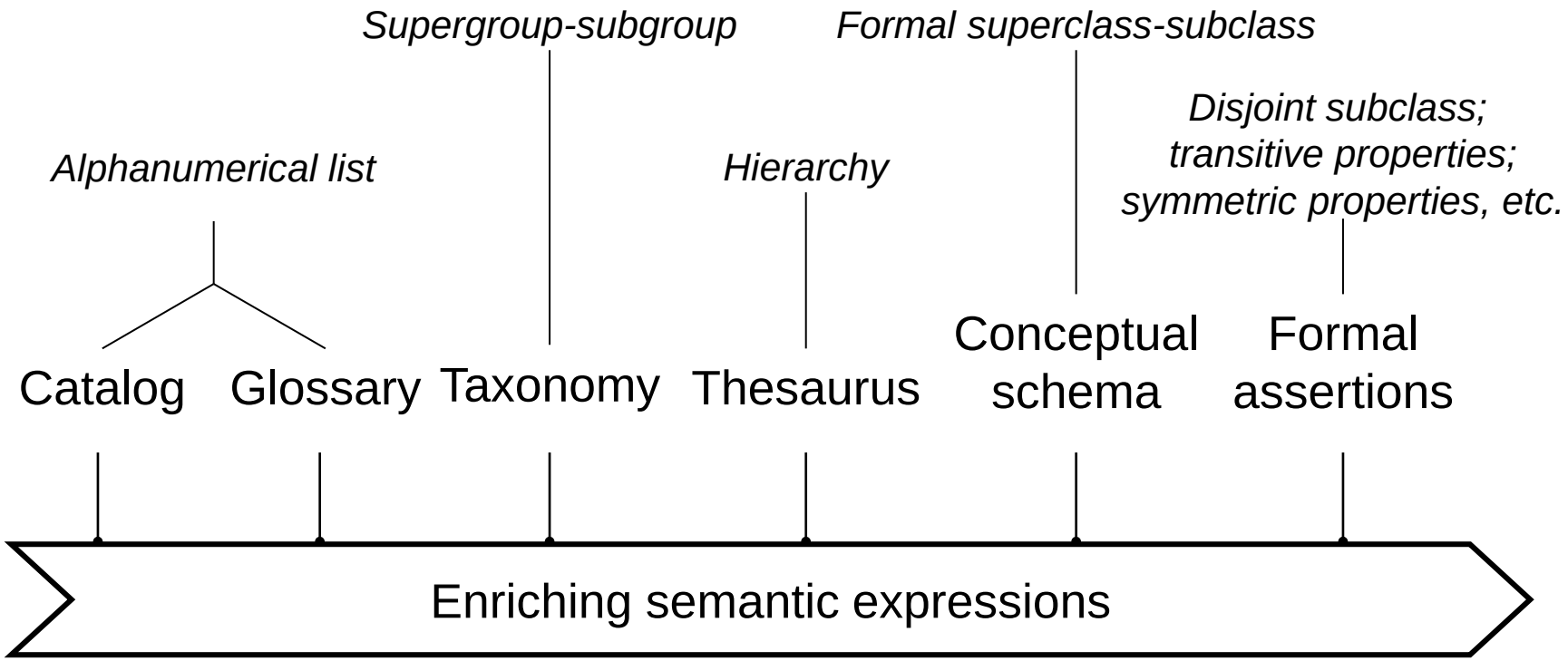


Aristotle
(384 – 322 BCE)

I Ching (Book of Changes)
(c. 450 – 250 BCE)

- For the Semantic Web purpose
  - An ontology is the specification of a shared conceptualization of a domain

An Ontology Spectrum



(Ma et al., 2010; adapted from Welty, 2002; McGuinness, 2003; Obrst, 2003; Uschold and Gruninger, 2004; Borgo et al., 2005)

Querying RDF Data

- Query Languages such as SPARQL
  - Most forms of the query languages contain a set of triple patterns
  - Triple patterns are like RDF triples except that each of the subject, predicate and object may be a variable

```
<Mozart, composed, The Magic Flute>
<Mozart, isA, Musician>
<The Magic Flute, isA, Opera>
```

Data

- ?x, ?y are variables
- ?x composed ?y represents a <subject, predicate, object> triple

```
SELECT   ?x   ?y
WHERE
{
  ?x composed ?y .
}
```

Query

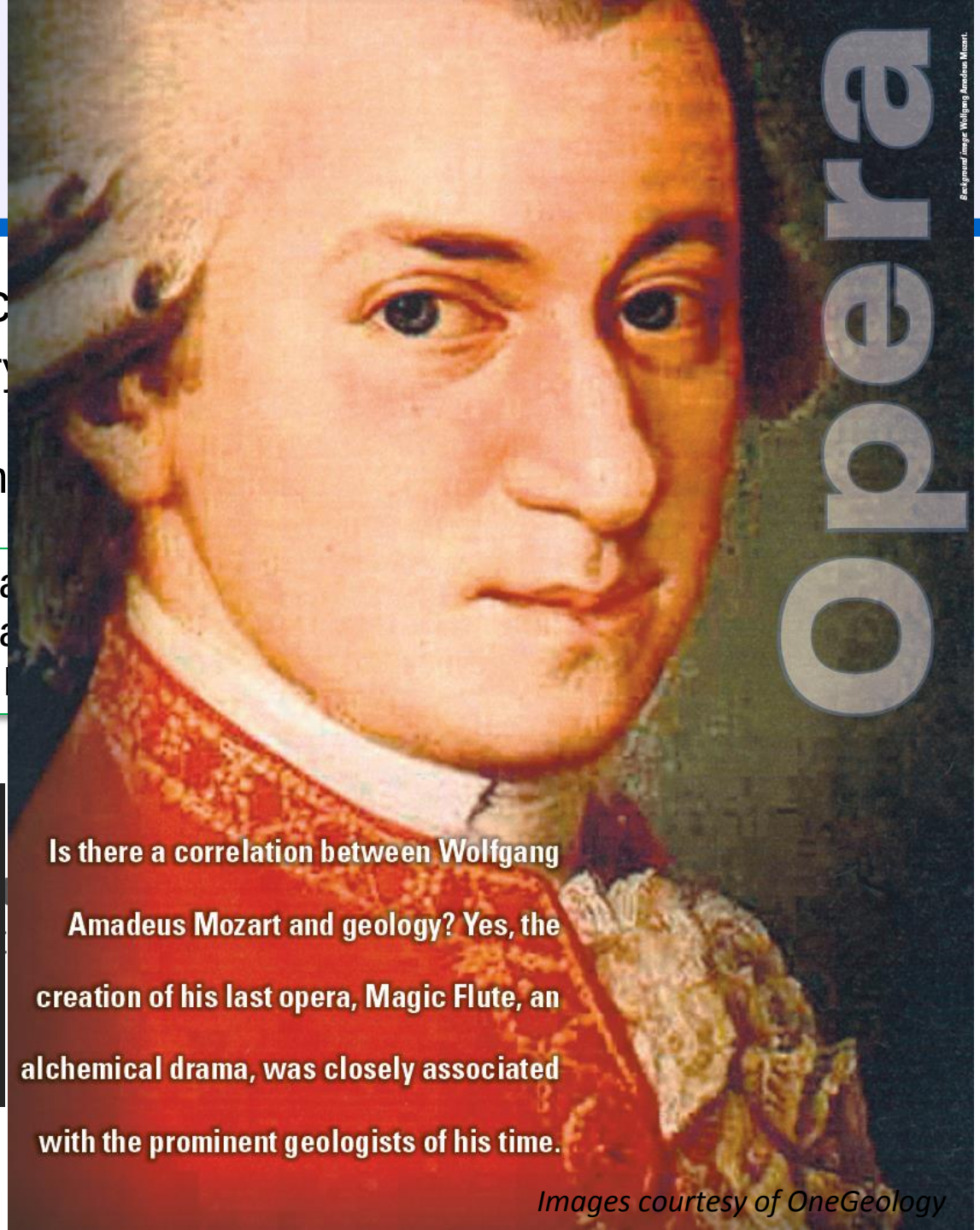
Mozart	The Magic Flute

Result

- Query Languages suc
  - Most forms of the quer
  - Triple patterns are like
    predicate and object m

<Moza
<Moza
<The

Correlation between Mozart and Geology?

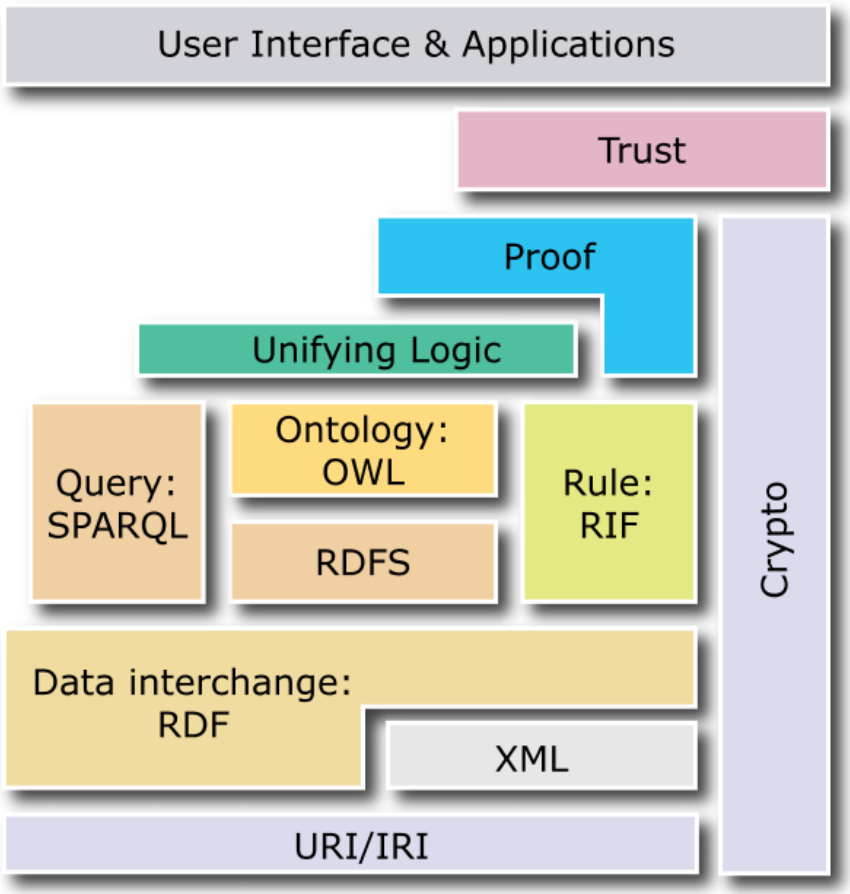


Images courtesy of OneGeology

A Vision of The Semantic Web

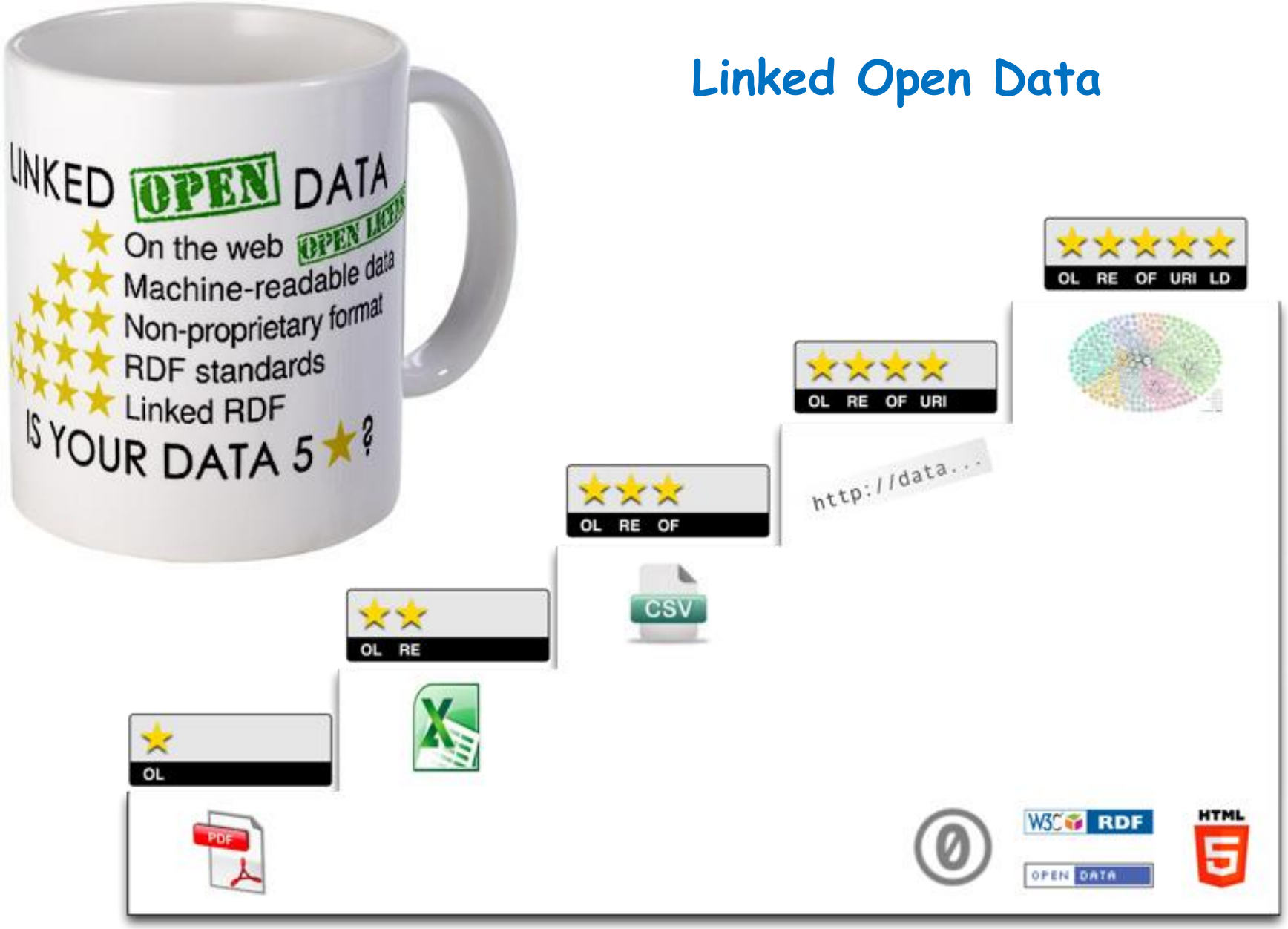
Machine-processable, global Web standards:

- Assigning unambiguous identifiers (URI)
- Expressing data, including metadata (RDF)
- Capturing ontologies (OWL)
- Query, rules, transformations, deployment, application spaces, logic, proof, trust



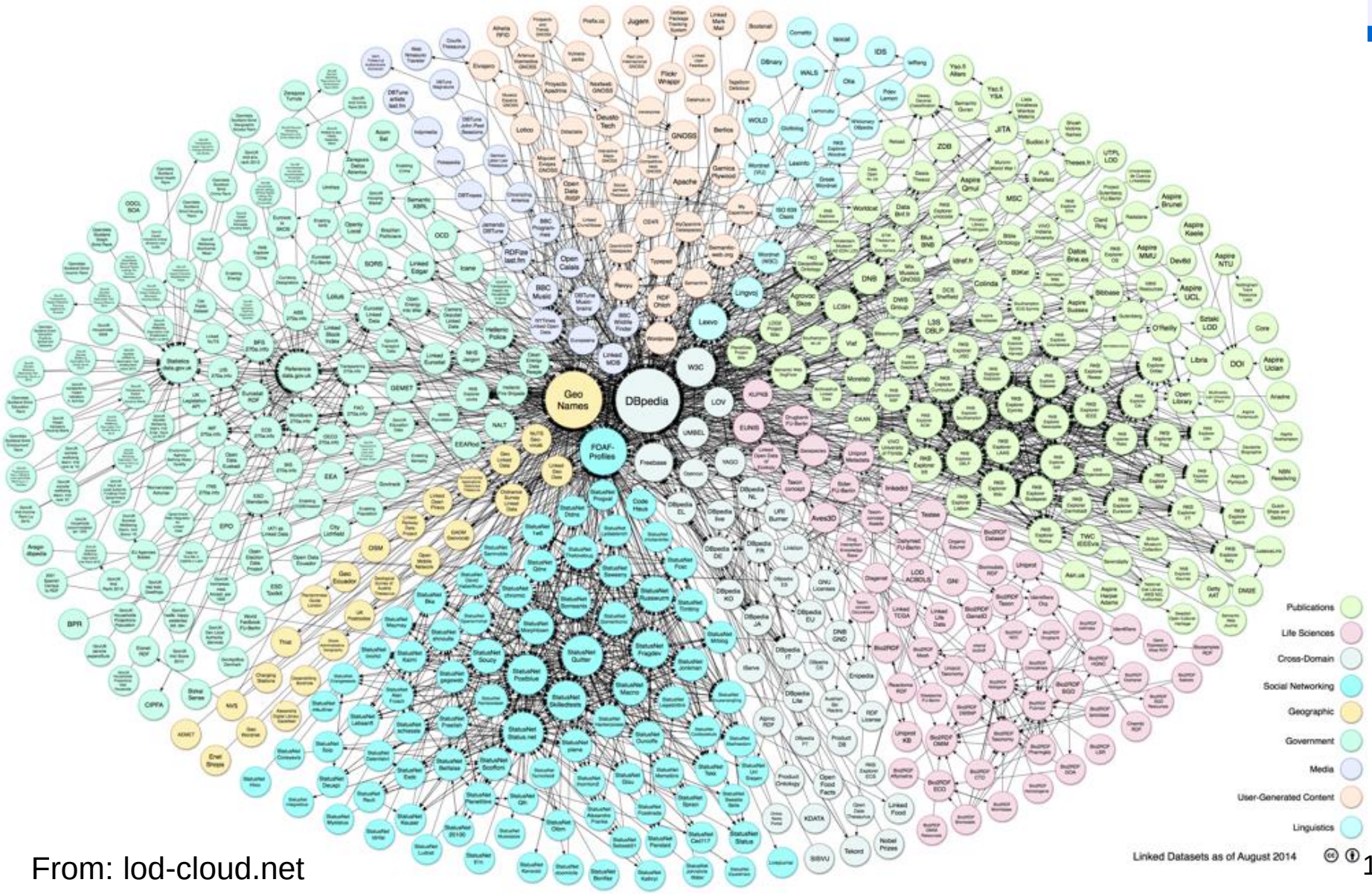
Bratt, 2006, *Emerging Web Technologies to Watch*

Linked Open Data



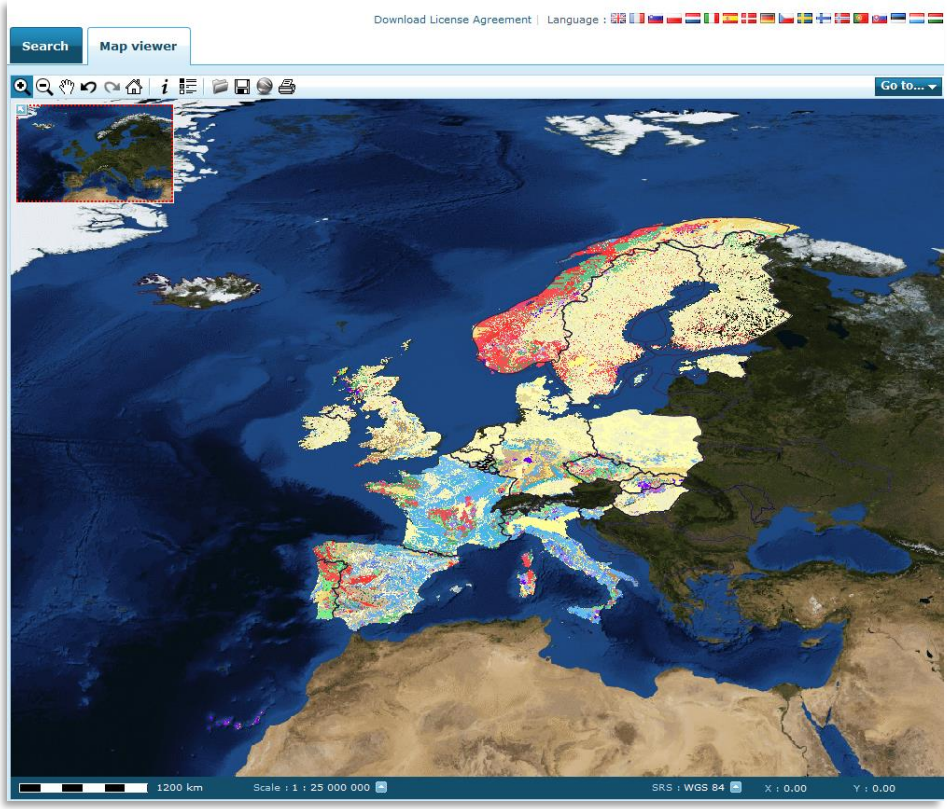
(Berners-Lee, 2006) (Image sources: w3c.org and 5stardata.info)

The Linking Open Data Cloud (Aug. 2014)



From: lod-cloud.net

Recent Works in Geoscience



OneGeology-Europe

- 20 European nations providing national geologic maps at scale ~1: 1M
- Harmonized geological terms and map legends
- Multilingual labels in 18 languages
- Central portal for data browsing/query among distributed data sources

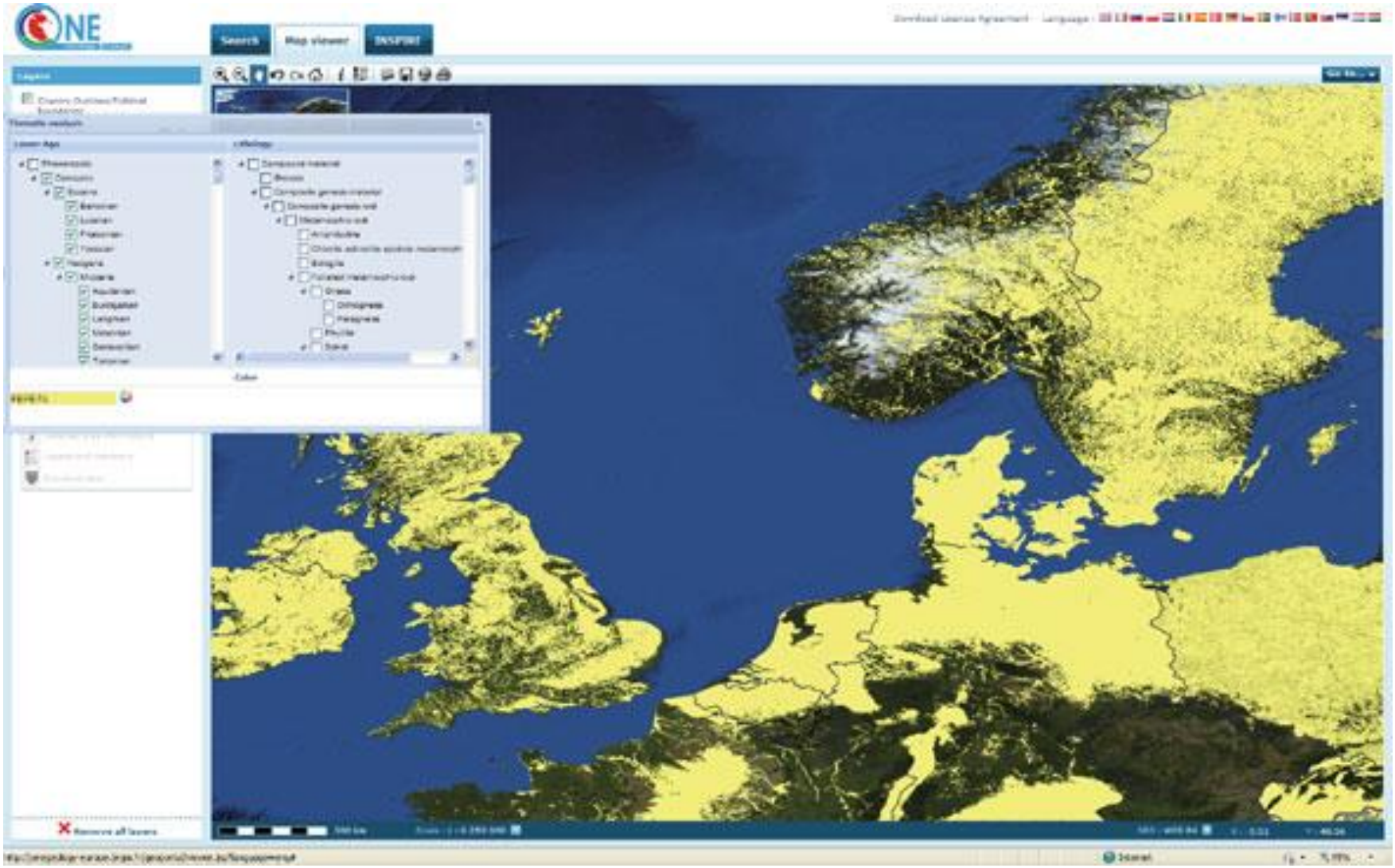
http://www.onegeology-europe.org

Co-funded by the European Union

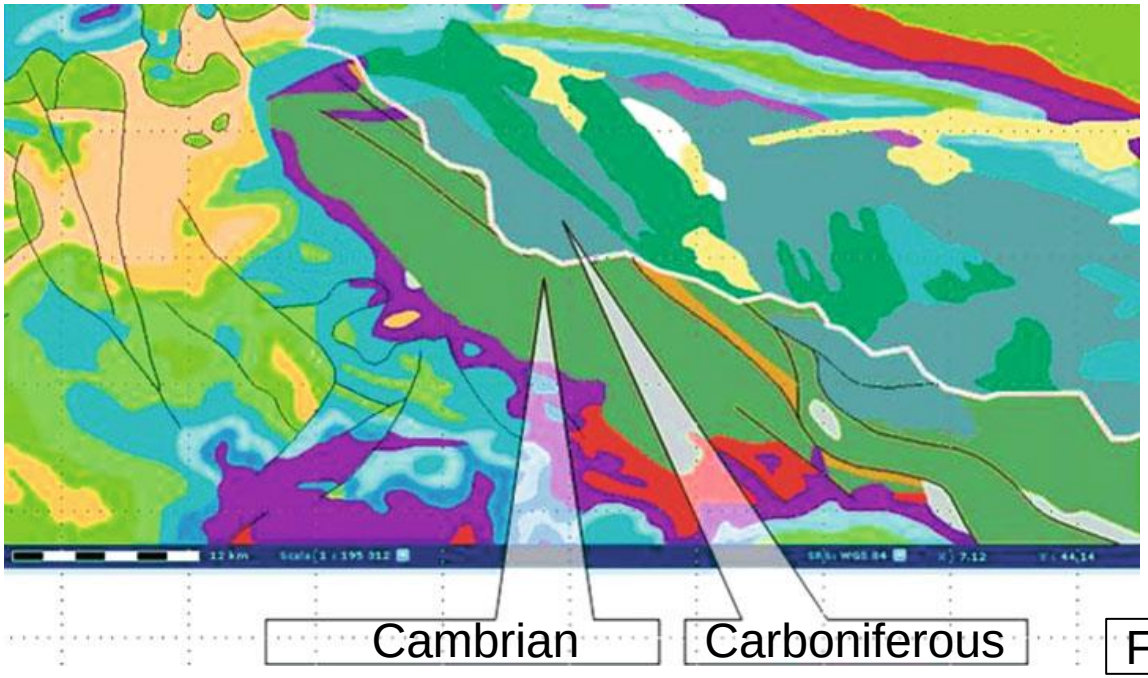
A contribution to INSPIRE

Federated query

Result of geologic units with age ‘Cenozoic - from 66 million years to today’



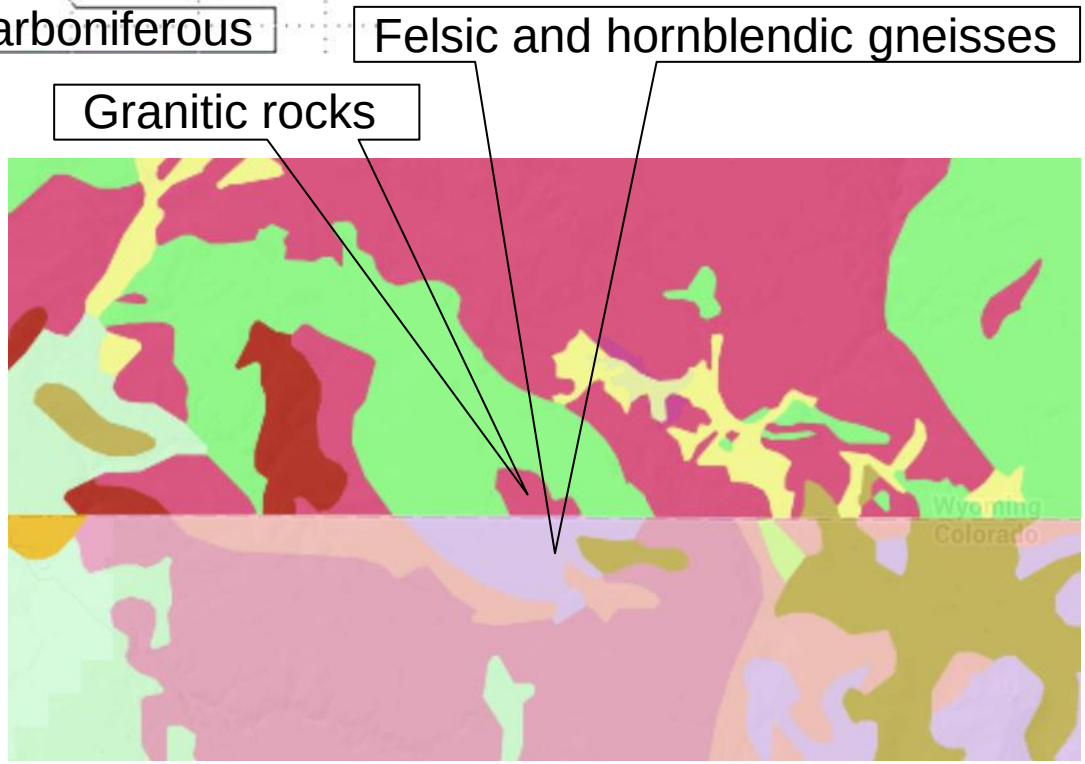
http://onegeology-europe.brgm.fr/geoportal/viewer.jsp



Italy/France near Cuneo/Colmar

(Asch et al., 2012)

There are still works to be done….



Wyoming/Colorado

(Ma et al., 2014)

Distributed datasets: Mismatches of geological units across political boundaries

(Base map courtesy: OneGeology-Europe and USGS)

Data Interoperability:
"Data should be discoverable, accessible, decodable, understandable and usable, and data sharing should be legal and ethical for all participants."

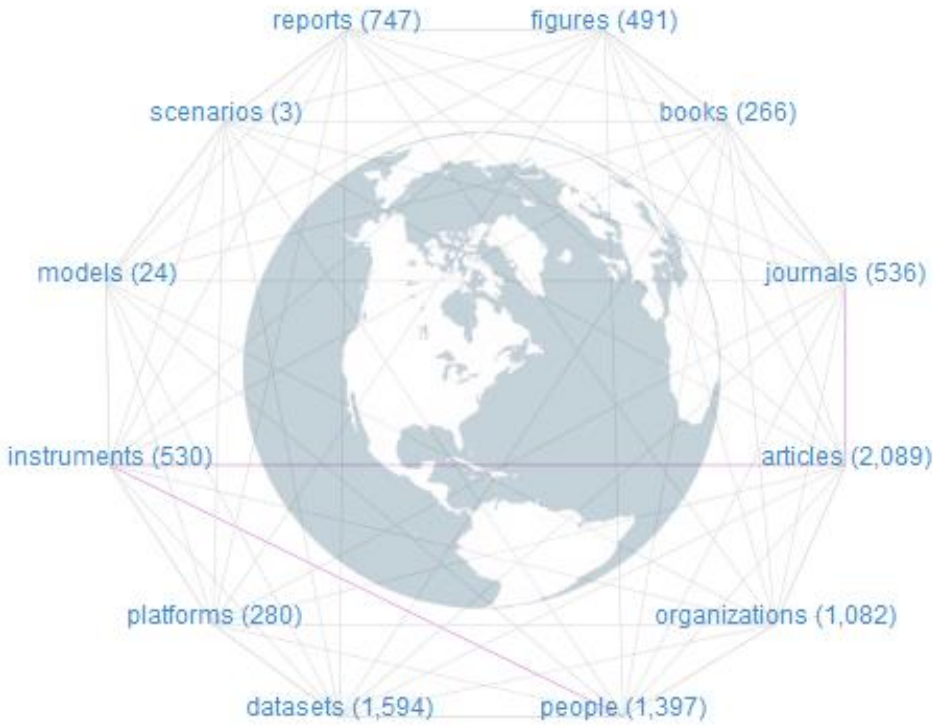
Ma et al., 2011, *nGeo*

Original image from: http://ehna.org

http://data.globalchange.gov

Global Change Information System

Providing structured global change information.

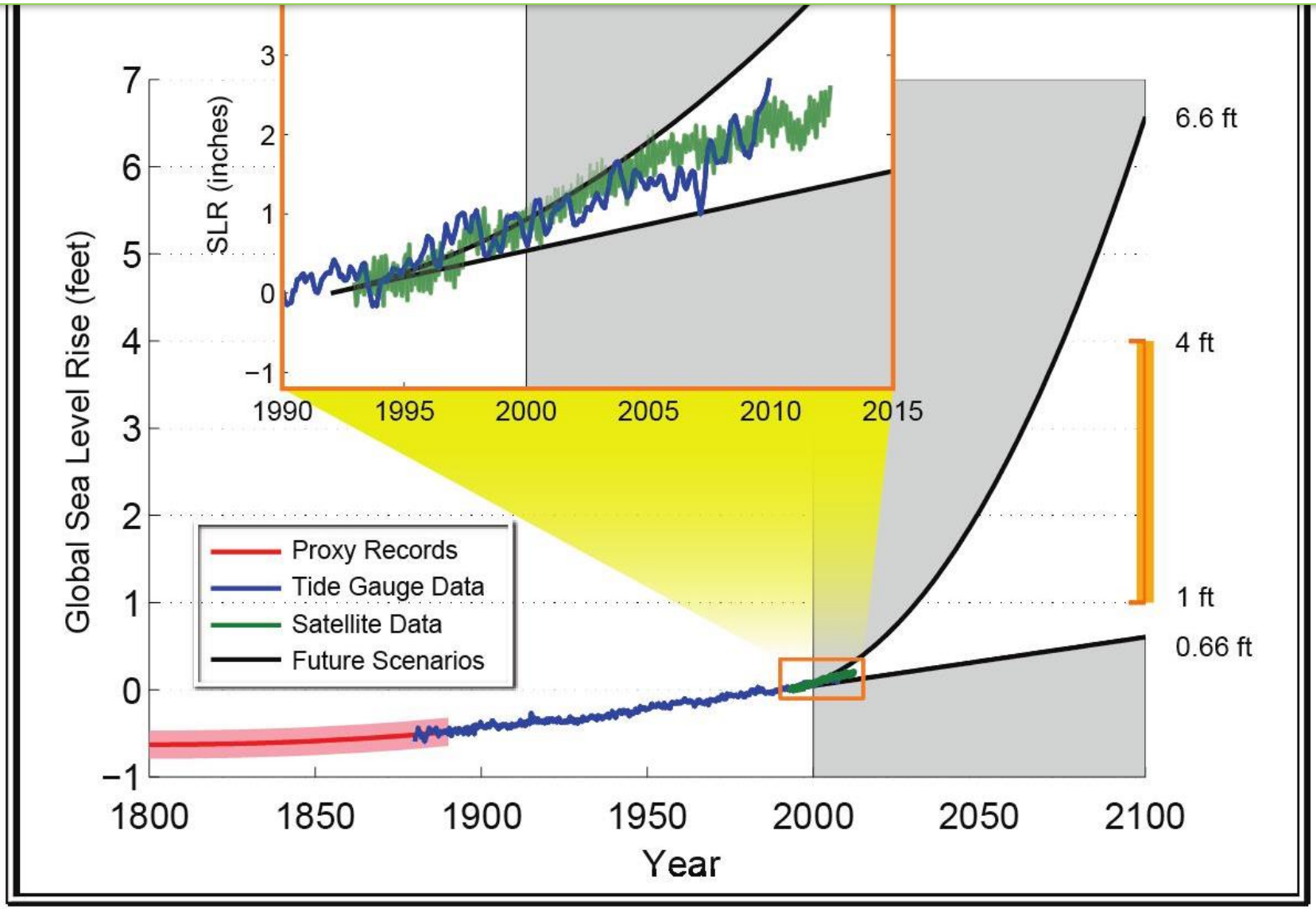


JSON YAML

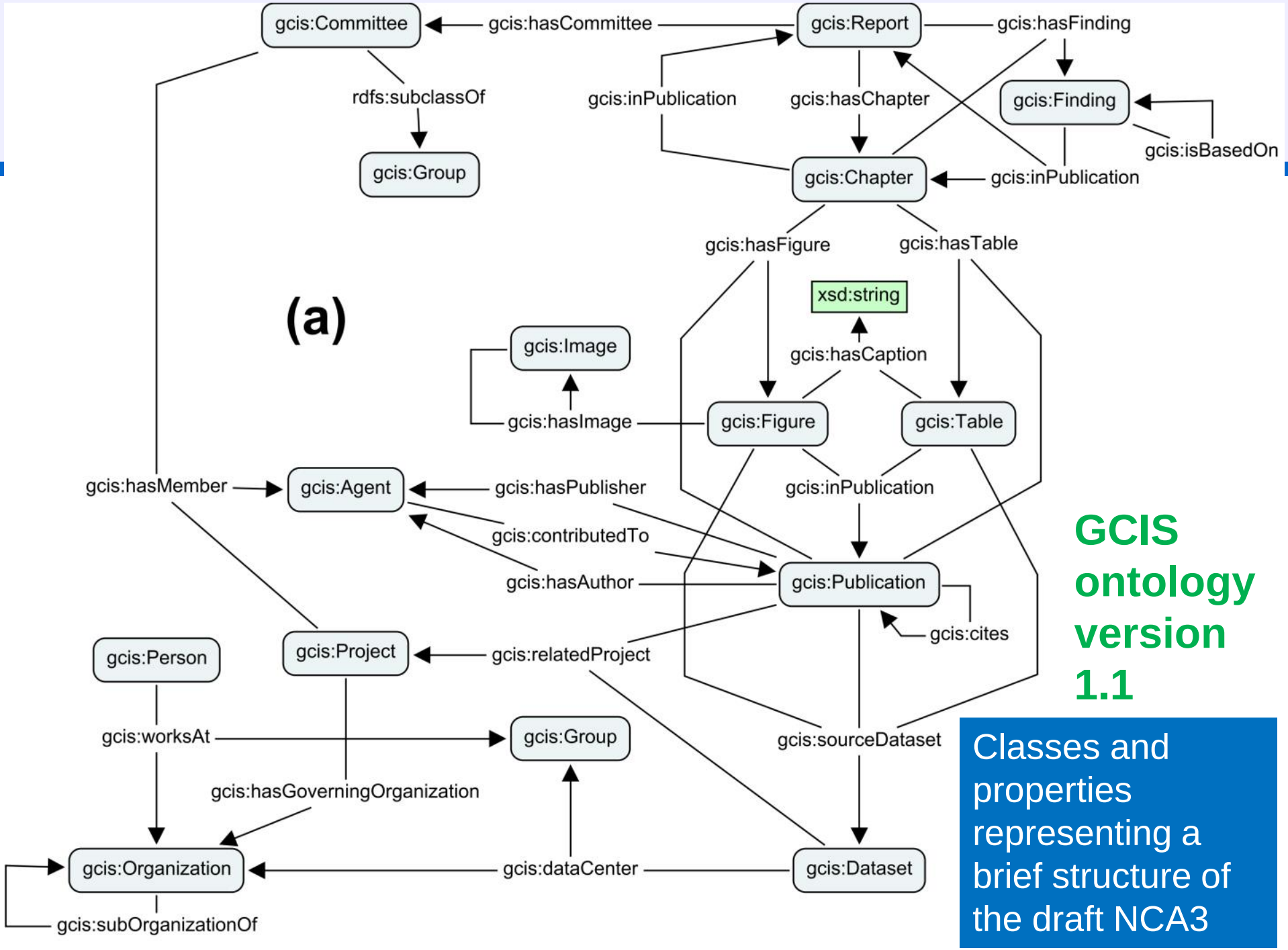
Featured report : The Third National Climate Assessment

An example question of provenance tracing:

What are the NASA contributions to Figure 1.2 in the draft NCA3?

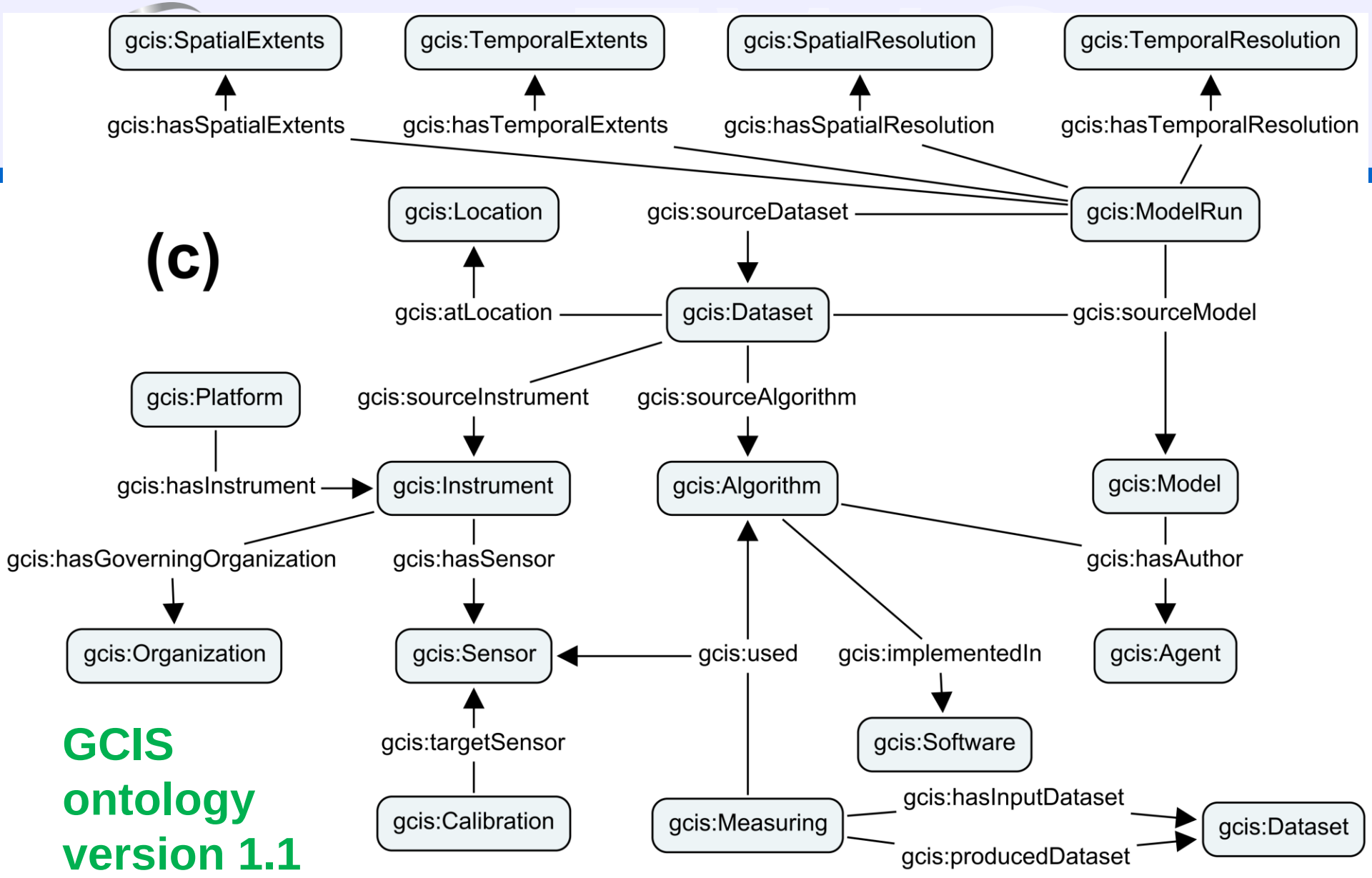


"*Figure 1.2: Sea Level Rise: Past, Present, and Future*" in draft NCA3



GCIS ontology version 1.1

Classes and properties representing a brief structure of the draft NCA3



GCIS ontology version 1.1

Classes and properties about sensors, instruments, platforms, and algorithms, etc. that datasets are generated from

Provenance Documentation:
"Linking a range of observations and model outputs, research activities, people and organizations involved in the production of scientific findings with the supporting data sets and methods used to generate them"

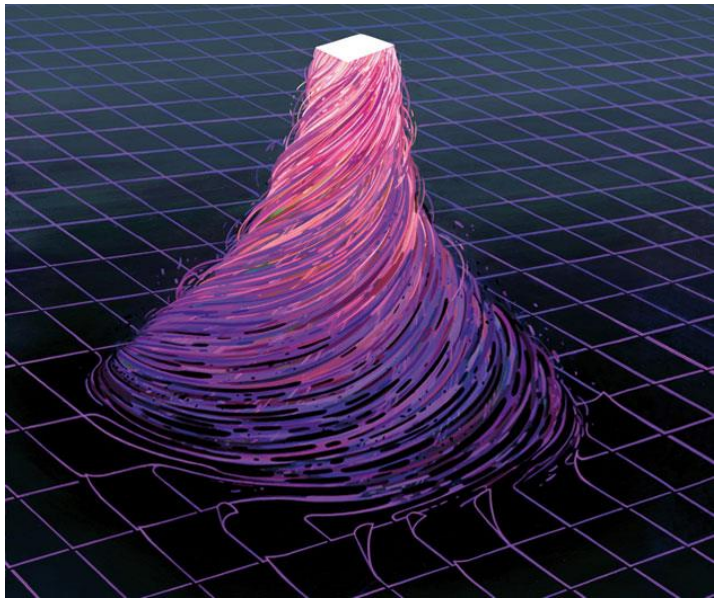


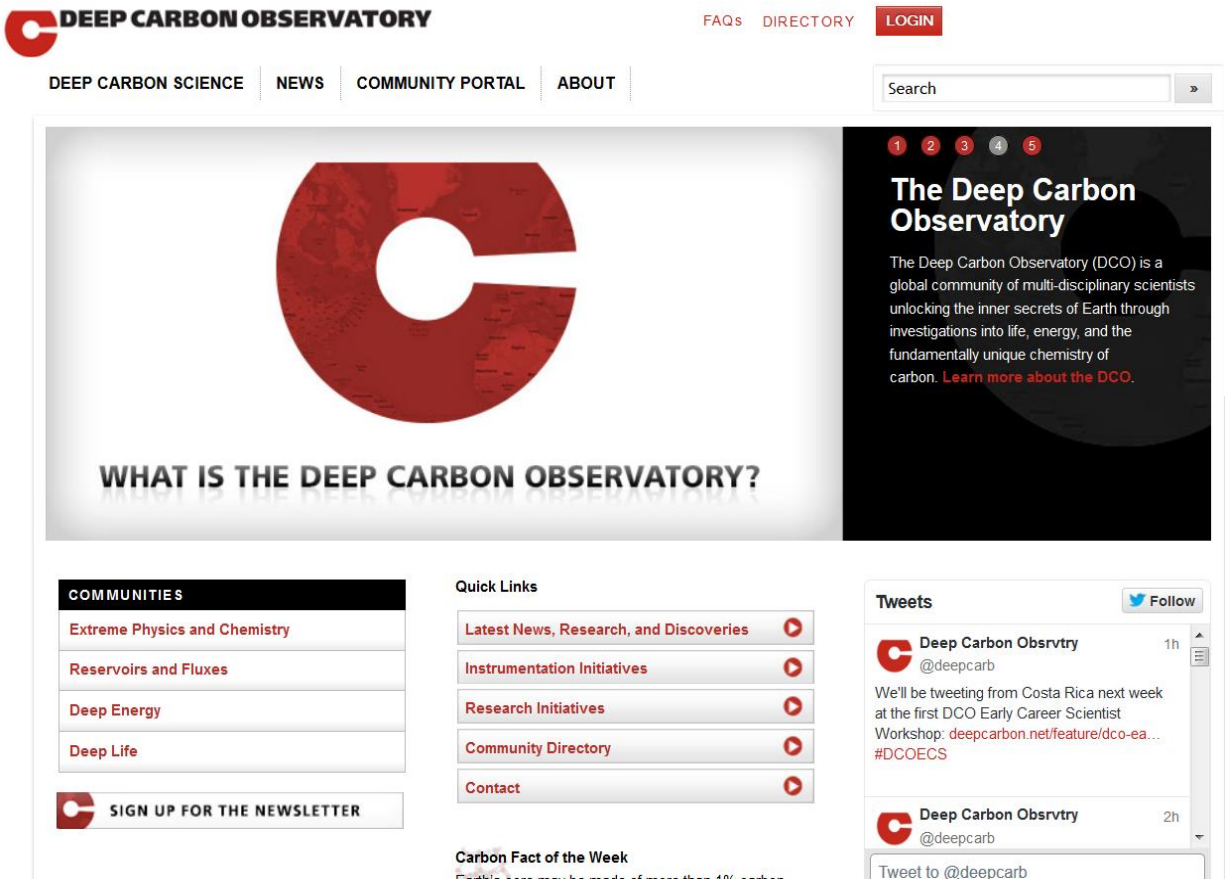
Image from nature.com

Ma et al., 2014, *nClimate*

United States
Global Change
Research Program

DEEP CARBON OBSERVATORY
DATA SCIENCE

Deep Carbon Virtual Observatory



Deep Carbon Observatory (2009-2019)

- Deep Energy
- Deep Life
- Extreme Physics and Chemistry
- Reservoirs and Fluxes

DCVO: A cyber-enabled platform for linked science

http://deepcarbon.net

ALFRED P. SLOAN FOUNDATION

Fox, 2014

Deep Carbon Virtual Observatory

- A vision of the DCVO:
  - A conceptual model of the interplay between data, people, publication, instruments, models, organizations, etc.
  - Identify, annotate and link all key entities, agents and activities
  - A repository for datasets and associated metadata
  - Unique and powerful data and metadata visualization for dissemination of information
  - Collaboration tools for scientific efforts
  - An integrated portal for diverse content and applications

http://deepcarbon.net

Fox et al., 2014

Deep Time Data Infrastructure (2015-2025)

Studying the co-evolution of geosphere and biosphere

Vast amounts of data related to planetary evolution through deep time:

- Mineralogy and Petrology
- Paleobiology and Paleontology
- Paleotectonics and Paleomagnetism
- Geochemistry and Geochronology
- Genomics and Proteomics

…

Deep Time Data Workshop at AGU 2014

Short-term goal (2015-2017):

Develop, curate, and integrate diverse data resources to focus on our planet's changing near-surface oxidation state and the rise of oxygen through deep time

http://www.wmkeck.org/grant-programs/research/medical-research-grant-abstracts/science-and-engineering-2014

Deep Time Data Infrastructure (2015-2025)

Studying the co-evolution of geo- and biospheres

Vast amounts of data related to planetary evolution through deep time:

- Mineralo
- Paleobio
- Paleotec
- Geocher
- Genomic

…

Hysted, G., Downs, R.T., and Hazen, R.M. (2015) Mineral frequency distribution data conform to a LNRE model: Prediction of Earth's "missing" minerals. ***Mathematical Geosciences***, *in press*.

Workshop at AGU 2014

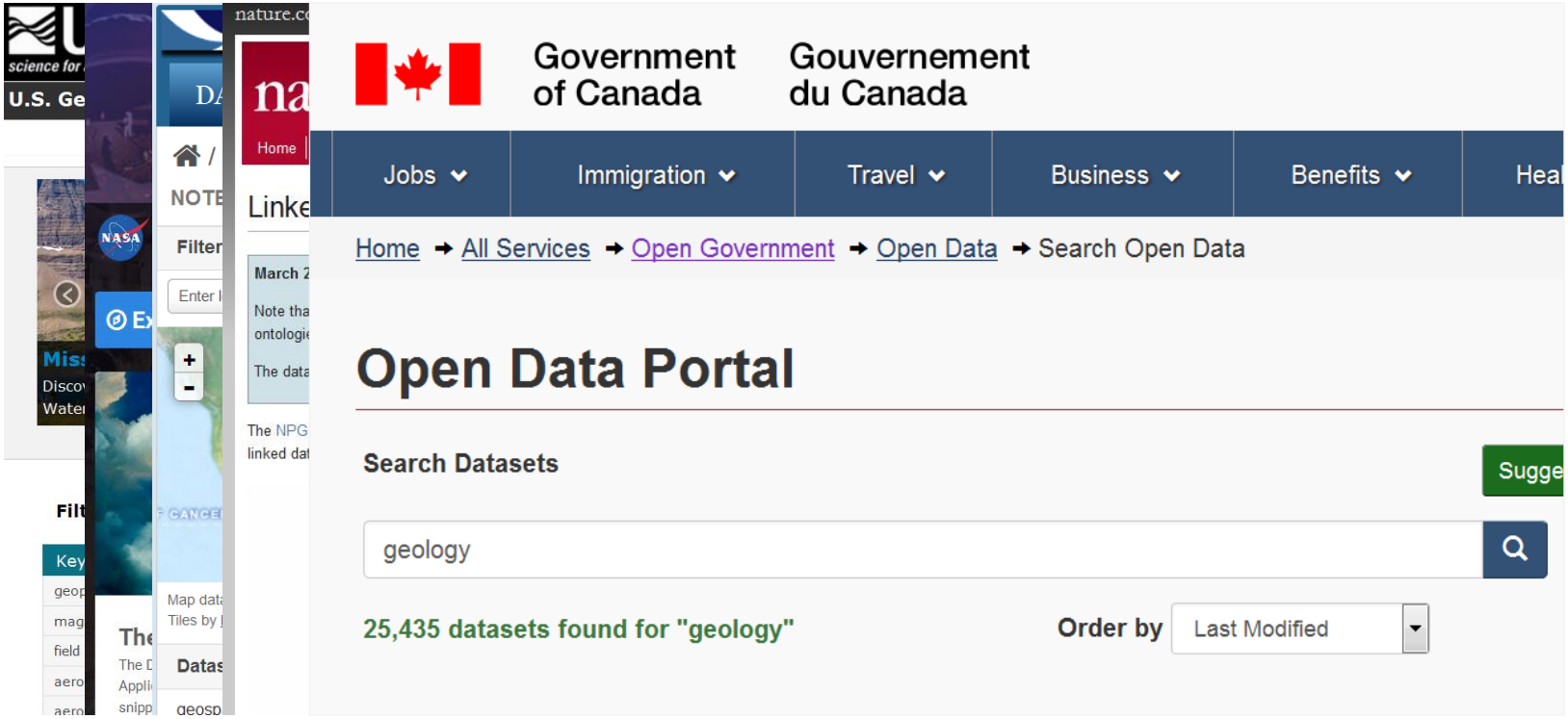
Short-term goal (2015-2017):

Develop, curate, and integrate diverse data resources to focus on our planet's changing near-surface oxidation state and the rise of oxygen through deep time

http://www.wmkeck.org/grant-programs/research/medical-research-grant-abstracts/science-and-engineering-2014

Exploring the Web of Data

- Geoscience vocabularies and ontologies are increasingly created and used
  - Concept recognition, comparison and interlinking will improve the quality of data integration



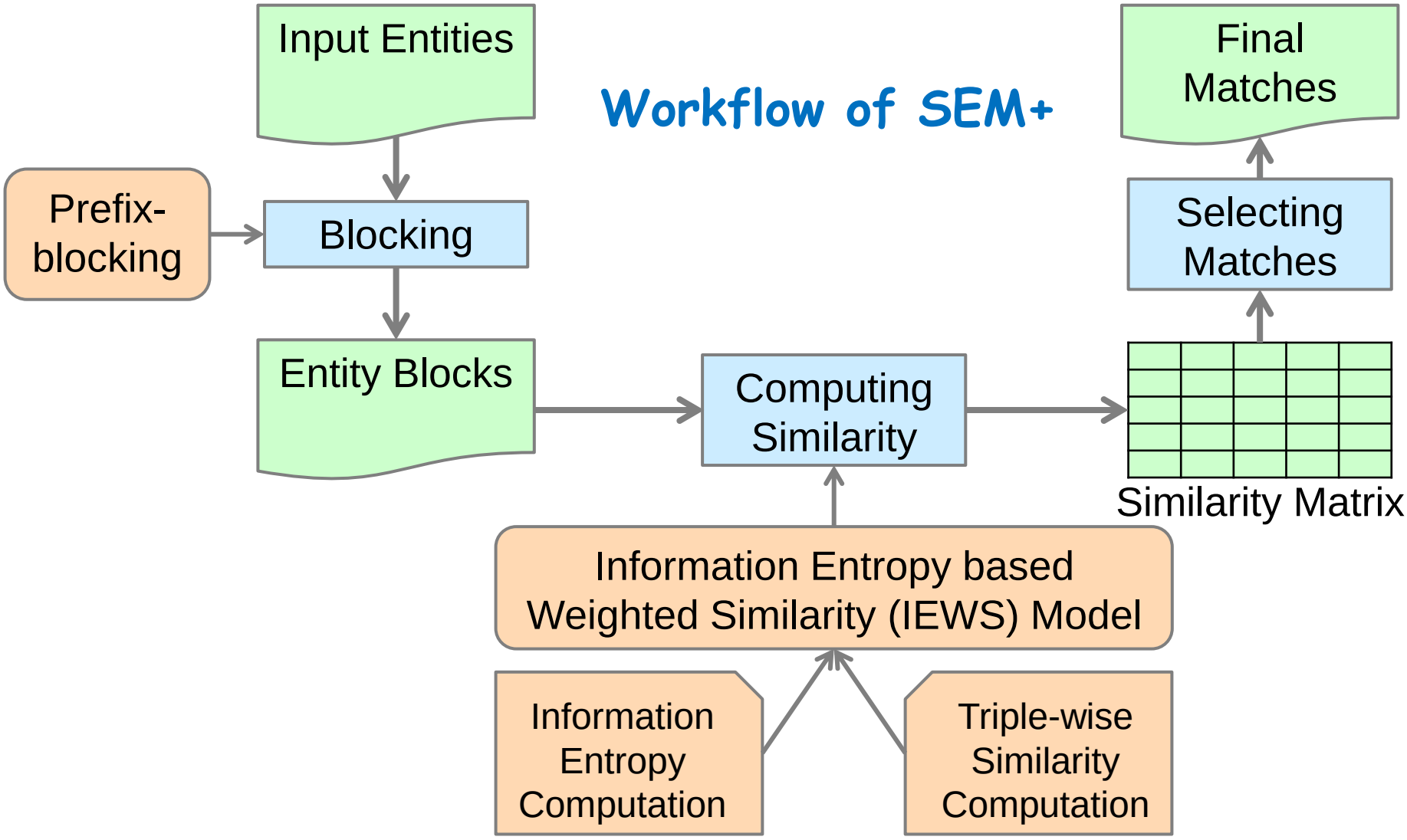
Exploring the Web of Data

- Geoscience vocabularies and ontologies are increasingly created and used
  - Concept recognition, comparison and inter-linking will improve the quality of data integration

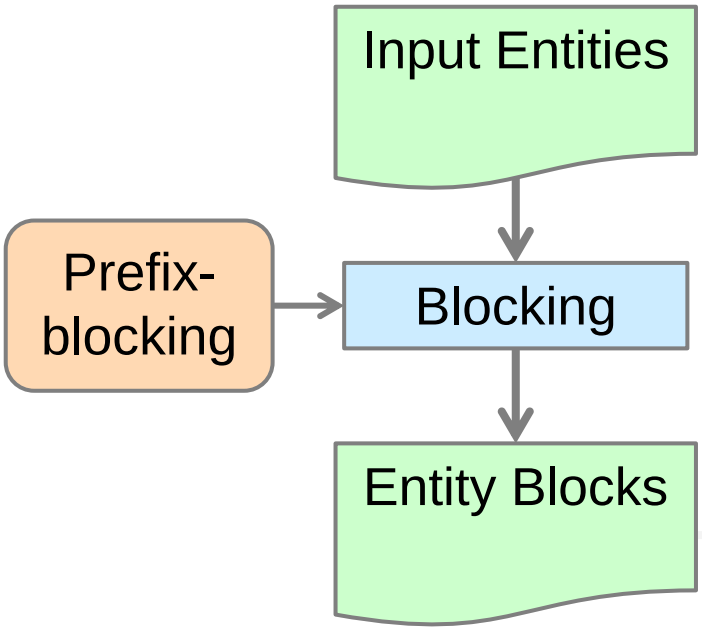
- SEM+: a tool for concept mapping in geoscience
  - SEM: Similarity-based Entity Matching
  - Compute semantic similarity between concepts
  - Suggest possible linking

Zheng et al., 2015, *ESIn*

SEM+: Similarity-based Entity Matching



Blocking Algorithm

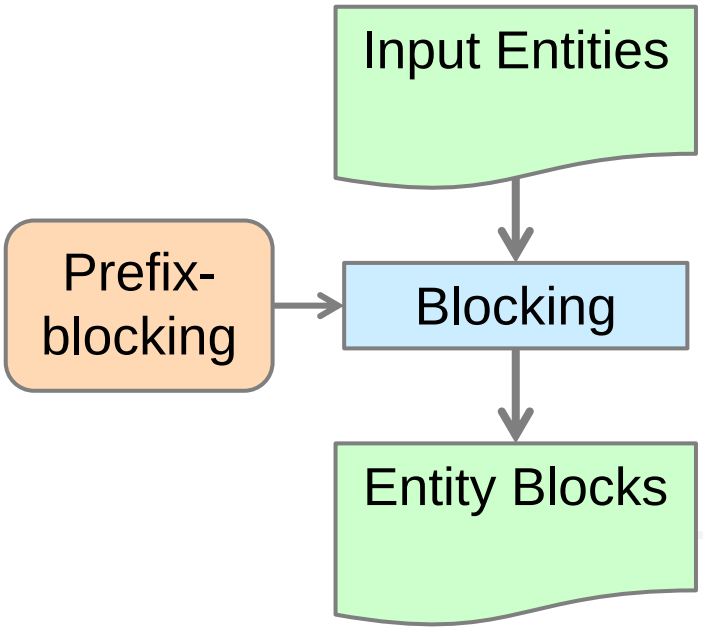


- Input: two or more large sets of concepts

An example concept

```
isc:Archean
    rdf:type gts:GeochronologicEra , skos:Concept ;
    rdfs:comment "younger bound-2500.0"@en , "older bound-4000.0"@en ;
    rdfs:label "Archean Eon"@en ;
    gts:rank <http://resource.geosciml.org/ontology/timescale/rank/Eon> ;
    … …
```

Prefix: Rare Keywords

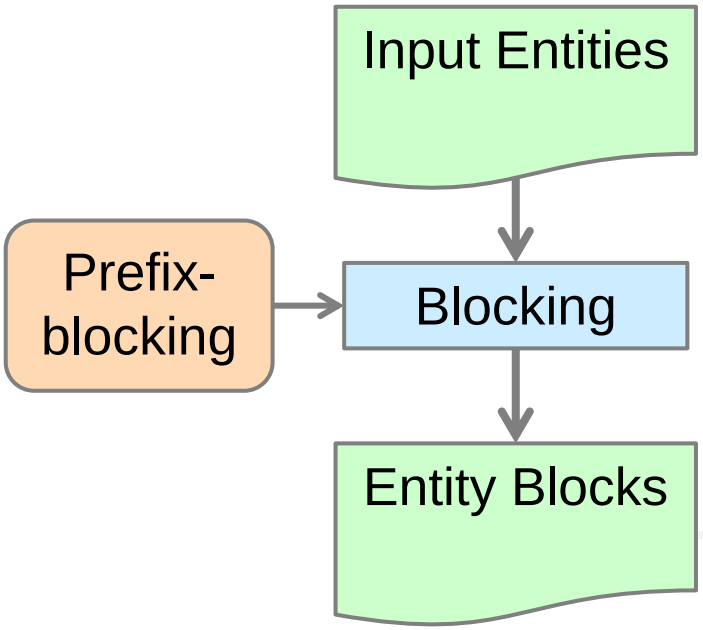


- Input: two or more large sets of concepts

- Blocking: group similar concepts

- Efficiency: reduce number of concept pairs
- Grouping concepts based on keywords in their literal descriptions
- Intuition: Concepts that share more rare keywords (prefix) are more likely to be similar
- Prefix-blocking: 'prefixes' are keywords that belong to the least number of concepts
- The final similarity computation will only apply to concepts in the same block

Concept Blocks



- Input: two or more large sets of concepts
- Blocking: group similar concepts
- Output: concept blocks

4 concepts & their keywords

w = {A, B, C, E, K, L}

x = {C,D, E, L}

y = {B, K, E, L}

z = {A, B, L}

Prefix-blocking

l_w: size of a block

l_b: blocking parameter

If $l_w > l_b$, remove that block

Result blocks

l_b=2, *then:*

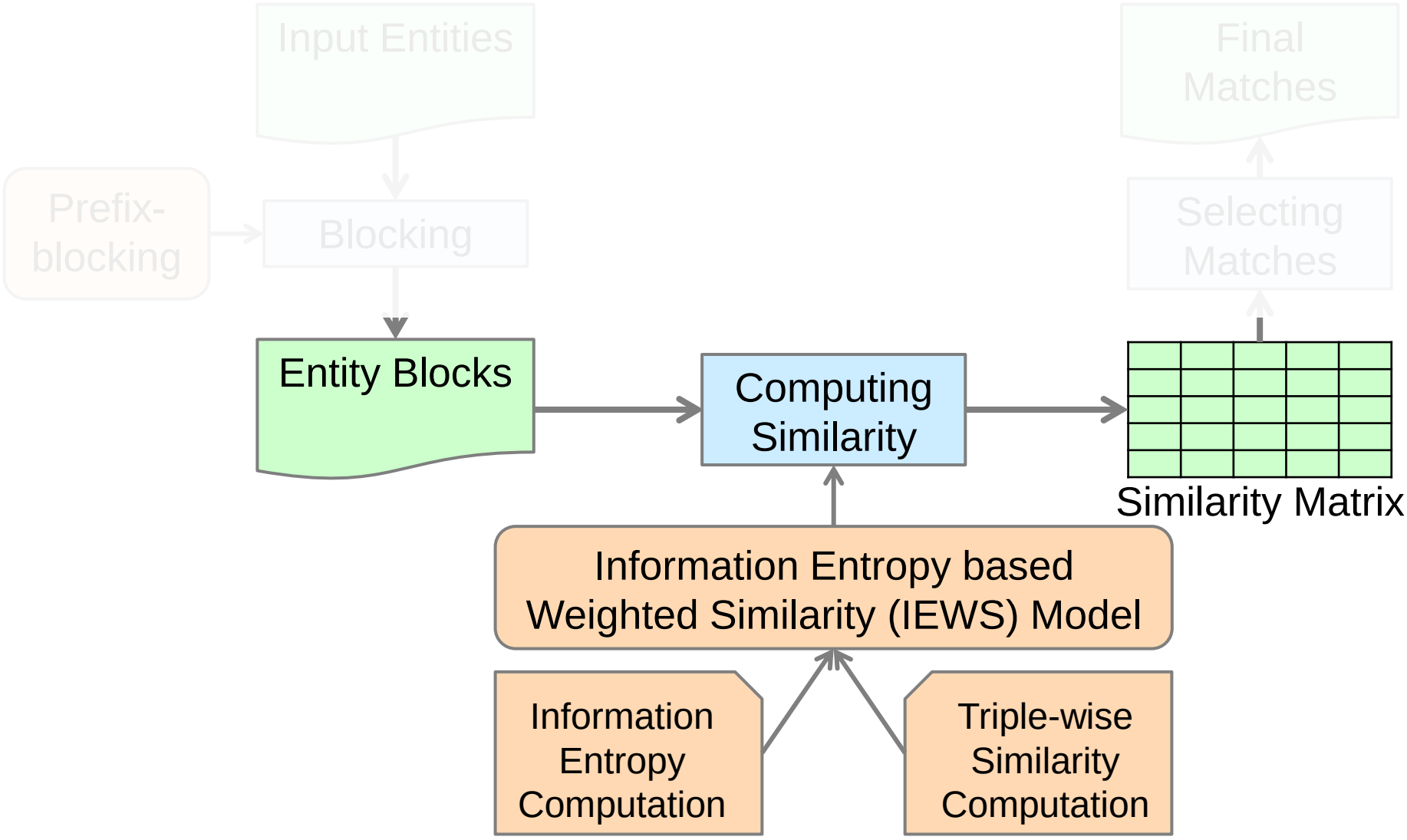
A: {w, z}

C: {w, x}

D: {x}

K: {w, y}

IEWS Model


Input Entities
Final Matches
Prefix-blocking
Blocking
Selecting Matches
Entity Blocks
Computing Similarity
Similarity Matrix
Information Entropy based Weighted Similarity (IEWS) Model
Information Entropy Computation
Triple-wise Similarity Computation

Triple-wise Similarity

- Similarity between two concepts *c* and *c'*
  - Similarity between triples describing the two concepts
  - Triple-wise (*pv*) similarity: Sim^{pv}
- A challenge: property mapping

Example: property mapping

_:Boston rdfs:type _:t1
_:t1 rdfs:label 'City'
is same as
_:Boston _:category 'City'.

$$Sim^{pv}(pv, pv') = \begin{cases} 1.\ Sim^{l}(v, v')\ if\ v\ and\ v'\ are\ both\ literal \\ 2.\ Sim^{F}(v, v')\ if\ v\ and\ v'\ are\ both\ URI \\ 3.\ Get\ literal\ value\ and\ then\ use\ Sim^{l} \\ \quad if\ v\ is\ literal\ and\ v'\ is\ URI \end{cases}$$

Triple-wise Similarity Computation

Similarity between Two Triple Values

- Similarity between two triple values
- For case 1, compute similarity using Lin's method (Lin, 1998. *ICML*)
- For case 2, use another equation $Sim^F(c, c')$ recursively
  - Here we only traverse URIs to the depth of three
- For case 3, first extract literal value of *v'* and then use Sim^l

$$Sim^{pv}(pv, pv') = \begin{cases} 1.\ Sim^l(v, v')\ if\ v\ and\ v'\ are\ both\ literal \\ 2.\ Sim^F(v, v')\ if\ v\ and\ v'\ are\ both\ URI \\ 3.\ Get\ literal\ value\ and\ then\ use\ Sim^l \\ \quad if\ v\ is\ literal\ and\ v'\ is\ URI \end{cases}$$

Triple-wise Similarity Computation

Similarity between Two Concepts

- Similarity between two concepts *c* and *c'*
- Apply Jaccard similarity (Jaccard, 1912. *New Phytologis*t)
- $|PV_1|$: number of *pvs* in concept *c*
- $|PV_2|$: number of *pvs* in concept *c'*
- α and β: coefficients of variation on the similarity measure on *c* and *c'* unique description

$$Sim(c, c') = \frac{\sum Sim^{pv}}{\sum Sim^{pv} + \alpha(|PV_1| - \sum Sim^{pv}) + \beta(|PV_2| - \sum Sim^{pv})}$$

Triple-wise Similarity Computation

Information Entropy

- Information entropy: Quantified measure of uncertainty of information content (Shannon, 1948)
- The amount of information of a property can be quantified as information entropy

Example

Properties are not equally important for concept description. A tripe describing the Social Security Number is more important than a triple of name, to identify a person

- X: a property, with possible values $\{x_1, x_2, x_3, \ldots, x_n\}$
- $P(x_i)$: possibility of X obtaining each value
- Information Entropy of X:

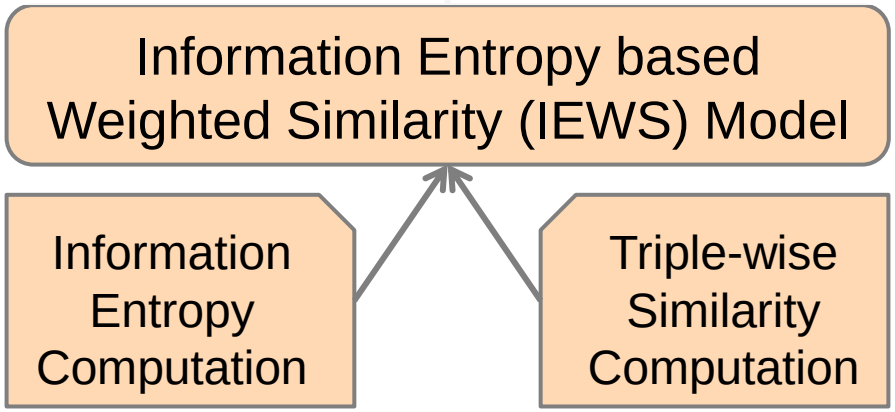
$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b (P(x_i))$$

Information Entropy Computation

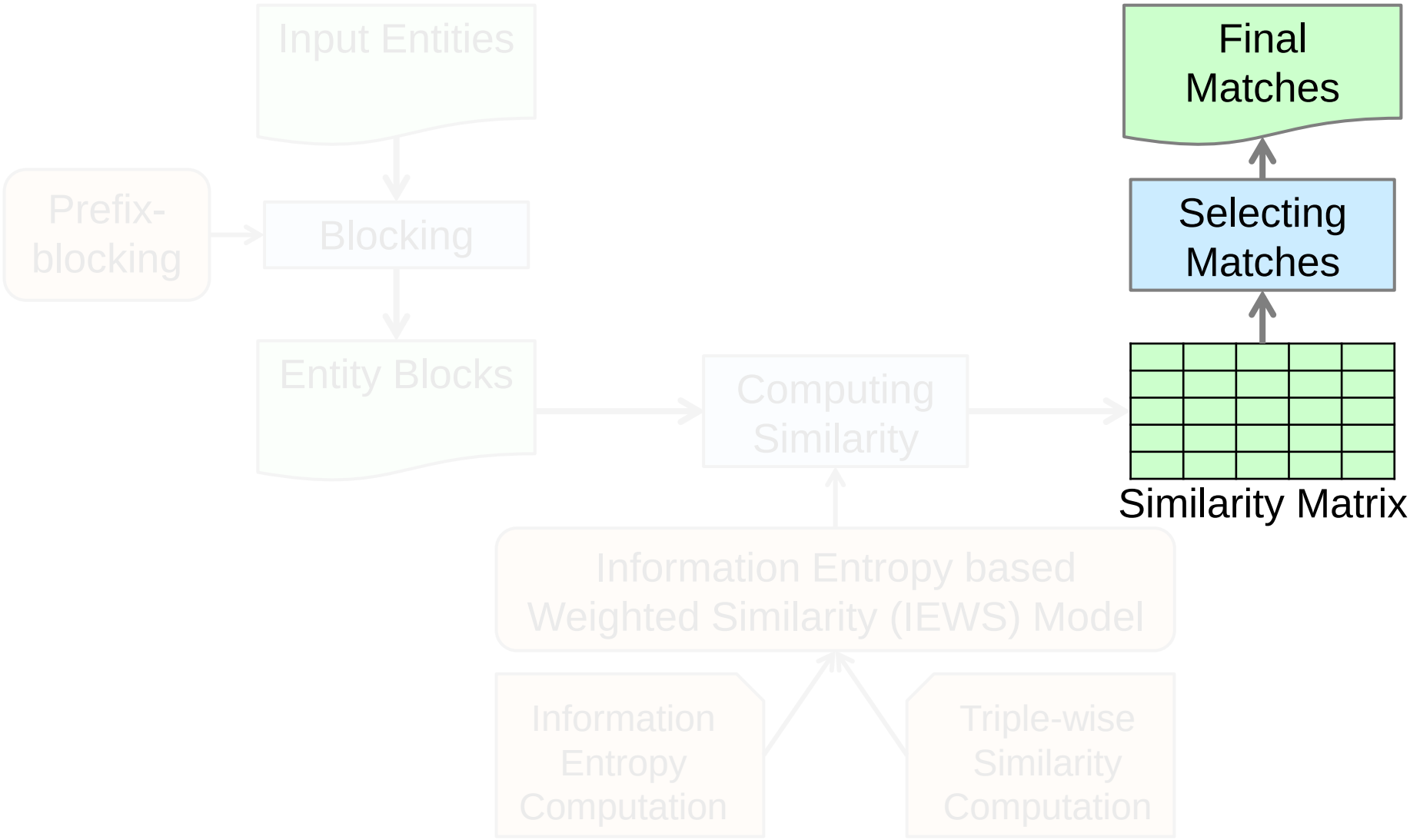
Information Entropy

- *P*: the set of properties in $\sum Sim^{pv}$
- *H(P)*: information entropy of the common descriptions

$$Sim^{F}(c,c') = H(P)\frac{\sum Sim^{pv}}{\sum Sim^{pv} + \alpha(|PV_1| - \sum Sim^{pv}) + \beta(|PV_2| - \sum Sim^{pv})}$$



Selecting Matches between Concepts



Summary

- eScience: the digital or electronic facilitation of science
- Semantic eScience
  - A virtuous circle between science and semantic technologies
  - Data driven + Knowledge driven?
- My understanding of Semantic eScience: AIR^3
  - **A**nyone can say **a**nything on **a**ny topic
  - **I**nteroperability, **i**nteractivity, **i**ntercreativity
  - The **r**ight information for the **r**ight person at the **r**ight time

Thanks for listening

max7@rpi.edu @MarshallXMa

- Backup slides

Earth Resource Form
Environmental Impact Value
Exploration Activity Type
Exploration Result
UNFC Value
Earth Resource Expression
Earth Resource Shape
Enduse Potential
Mineral Occurrence Type
Mining Activity Type
Processing Activity Type
Mining Waste Type Value
Commodity Code
Mineral Deposit Group
Mineral Deposit Type
Product Value

CGI Geoscience Terminology Workgroup

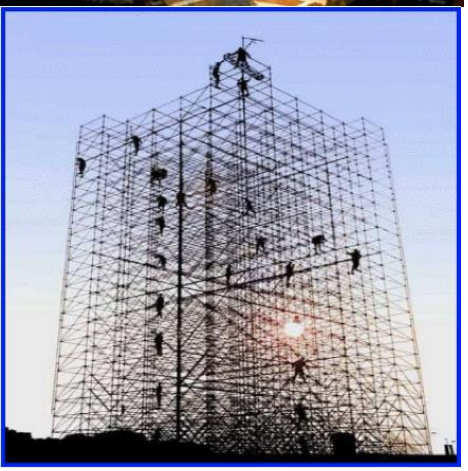
- Construct a collection of vocabularies for populating information interchange documents and enabling interoperability
- Provide labels for concepts, scope to various communities defined by language, science domain, or application domain

http://cgi-iugs.org/tech_collaboration/geoscience_terminology_working_group.html

Recently finished CGI vocabularies

Prior to 2005, we built systems; Now we build frameworks

- Rough definitions
  - Systems have very well-define entry and exit points. A user tends to know when they are using one. Options for extensions are limited and usually require engineering
  - Frameworks have many entry and use points. Users often do not know when they are using one. Extension points are part of the design
  - Platforms are built on frameworks



(Fox, 2014)

Semantic eScience

- Artificial Intelligence accelerates scientific discovery
  - Data search, synthesis and hypothesis representation
  - Data analysis: reasoning with models of the data

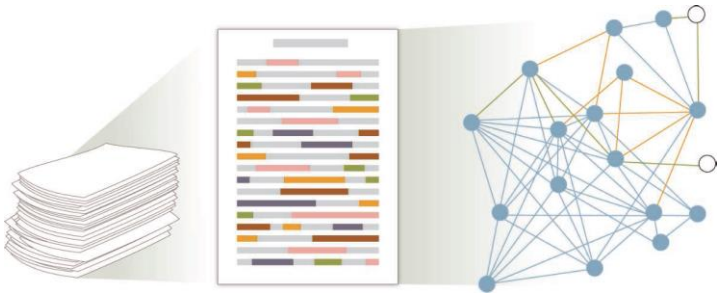


Image from science.com

A state-of-the-art example:
Hanalyzer (high-throughput analyzer)

- Uses natural language processing to automatically extract a semantic network from all PubMed papers relevant to a scientist
- Uses Semantic Web technology to integrate assertions from other biomedical sources
- Reasons about the network to find new correlations that suggest new genes to investigate

(Gil et al., 2014) (Leach et al., 2009)